Claims

1. A method for revalidating previously generated statistics for a query directed to one or more attributes of a relation, comprising

identifying in said query a selection criterion on said one or more attributes of said relation, and

revalidating a prior statistic generated for a prior different selection criterion on the same one or more attributes of said relation, based upon a measure of entropy of said one or more attributes of said relation.

2. The method of claim 1 wherein said prior statistic is revalidated if a measure of entropy of said one or more attributes of said relation is less than a predetermined threshold value.

3. The method of claim 1, further comprising generating a measure for the entropy of said one or more attributes of said relation, by the steps of

computing frequencies of different values for the one or more attributes in tuples of the relation, and

combining the measured frequencies into a measure of the entropy of the attributes.

4. The method of claim 3, wherein generating a measure for the entropy of said one or more attributes of said relation further comprises

collecting a sample of tuples of the relation, wherein frequencies of different values are computed for tuples in the sample.

5. The method of claim 3 wherein combining the measured frequencies comprises determining a number of distinct values for the one or more attributes, and converting the computed frequencies to probabilities by dividing the frequencies by number of distinct values.

6. The method of claim 5 wherein combining the measured frequencies further comprises forming a weighted sum of the computed probabilities.

7. A computer system implementing a relational database system and evaluating queries directed to said relational database, comprising

storage for said relational database, including a relation having a plurality of tuples including values for a plurality of attributes, and

computing circuitry performing query optimization and query execution upon said relational database, said query optimization including generating statistics for a query directed to one or more attributes of said relation, by identifying in said query a selection criterion on said one or more attributes of said relation, by revalidating a prior statistic generated for a prior different selection criterion on the same one or more attributes of said relation, based upon a measure of entropy of said one or more attributes of said relation.

8. A program product for implementing a relational database system and evaluating queries directed to said relational database, comprising

a relational database, including a relation having a plurality of tuples including values for a plurality of attributes, and

relational database software performing query optimization and query execution upon said relational database, said query optimization including generating statistics for a query directed to one or more attributes of said relation, by identifying in said query a selection criterion on said one or more attributes of said relation, by revalidating a prior statistic generated for a prior different selection criterion on the same one or more attributes of said relation, based upon a measure of entropy of said one or more attributes of said relation, and

a signal bearing media holding said relational database and relational database software.

9. The program product of claim 8 wherein the signal bearing media comprises transmission media.

10. The program product of claim 8 wherein the signal bearing media comprises recordable media.

11. A method for identifying a group of attributes of a relation for which a multi-dimensional index is to be formed, comprising

computing a correlation of attribute values within tuples of the relation, and

forming a multi-dimensional index for a group of attributes within tuples of the relation having a correlation of attribute values in excess of a threshold.

12. The method of claim 11, wherein computing a correlation of attribute values within tuples of the relation comprises collecting a sample of tuples of the relation, and computing correlation of attribute values within the sampled tuples.

13. The method of claim 11 wherein a correlation of attribute values is computed as an information gain for those attributes by comparing, for a common set of tuples, a sum of individual entropies of values of each attribute, to a joint entropy of the values of all attributes.

14. The method of claim 13, wherein a measure for the entropy of one or more attributes is generated by computing frequencies of different values for the one or more attributes in tuples of the relation, and combining the measured frequencies into a measure of the entropy of the one or more attributes.

15. The method of claim 14, wherein generating a measure for the entropy of said one or more attributes of said relation further comprises collecting a sample of tuples of the relation, wherein frequencies of different values are computed for tuples in the sample.

16. The method of claim 14 wherein combining the measured frequencies comprises determining a number of distinct values for the one or more attributes, and converting the computed frequencies to probabilities by dividing the frequencies by number of distinct values.

17. The method of claim 16 wherein combining the measured frequencies further comprises forming a weighted sum of the computed probabilities.

18. The method of claim 11 further comprising evaluating attribute groups found to have correlation to identify primary sources of correlation, by determining a mutual information gain by comparing information gain for a group of attributes, to the largest information gain of any sub-group of fewer of the same attributes.

19. The method of claim 18 wherein a multi-dimensional index is formed for an attribute group having information gain greater than a

threshold, if there is no larger attribute group including the same attributes having a mutual information gain greater than a threshold.

20. The method of claim 11 wherein correlation of attribute values is computed for all combinations of attributes of a relation, or alternatively by sampling a set of attribute groups and then evaluating other related groups of those found to have substantial correlation.

21. A computer system implementing a relational database system including indexes for said relational database, comprising

storage for said relational database, including a relation having a plurality of tuples including values for a plurality of attributes, and

computing circuitry performing query execution upon said relational database, and identifying a group of attributes of a relation for which a multi-dimensional index is to be formed, by computing a correlation of attribute values within tuples of the relation, and forming a multi-dimensional index for a group of attributes within tuples of the relation having a correlation of attribute values in excess of a threshold.

22. A program product for implementing a relational database system, comprising

a relational database, including a relation having a plurality of tuples including values for a plurality of attributes, relational database software performing query execution upon said relational database, and identifying a group of attributes of a relation for which a multi-dimensional index is to be formed, by computing a correlation of attribute values within tuples of the relation, and forming a multi-dimensional index for a group of attributes within tuples of the relation having a correlation of attribute values in excess of a threshold, and

a signal bearing media holding said relational database and relational database software.

23. The program product of claim 22 wherein the signal bearing media comprises transmission media.

24. The program product of claim 22 wherein the signal bearing media comprises recordable media.